Design of experiments with a ranking response: analysis of the result with the Mann-Whitney statistic

Maurice Pillet, Emmanuel Duclos, Magali Pralus

Design of experiments with a ranking response: analysis of the result with the Mann-Whitney statistic

Maurice Pillet[(1)], *Emmanuel Duclos*[(2)], *Magali Pralus*[(1)]
[(1)] University of Savoy - SYMME – Polytech'Savoie - FRANCE
[(2)] EDC Conseil Formation
maurice.pillet@univ-savoie.fr
educlos@free.fr
magali.pralus@univ-savoie.fr

Summary: Product quality very often depends on organoleptic characters that are difficult to measure. As examples, let us look at the visual aspect of a vehicle dashboard, the flavour of a product... During the process optimization, it's very difficult to use such responses to analyse an experimental design, because of the lack of information contained in this type of response and the problems of repeatability and reproducibility inherent in these characters. However, if it is not possible for an appraiser to provide a measure in a continuous scale, it is easier to compare various objects. In this article, we propose to use this classification to calculate a rank variable (Mann-Whitney statistic) which will be used as a numeric variable in order to exploit the results of an experimental design. Several strategies will be presented and illustrated with industrial examples
Keywords: Mann-Whitney statistic, sensory perception defects, sensory analysis, design of experiments

1 Introduction

1.1 Context

Customer perception of product quality is not only related to the functional aspect of the product but also to sensory characteristics. This is particularly true for luxury products, and industrialists from all branches of industry pay attention to organoleptic characteristics.
Today, quality improvement problems in sensory perception defects cases are a problem in many companies. The majority of sensory perceptions are very difficult to characterize using sensors and companies must trust the controllers' perception to evaluate the quality of the products. Methods of sensory analysis well described in [1][2][3][4][5] propose several approaches allowing the quantification of this type of characteristic particularly in the food industry.
However, in spite of the relevance of this type of description, human perception does not provide a continuous measurement. The lack of measurement is a true difficulty for quality improvement since the majority of statistical tools are inapplicable if there are not numerical results. In particular we think of design of experiments which are tools that cannot be ignored to improve quality.
If the expert is not able to produce a measurement, he/she is nevertheless often able to rank various products according to his/her perception. On the basis of such ranking, we propose in this article to transform this ranking into a numeric variable by using the Mann-Whitney statistic.

1.2 Suggested method

The suggested method can be divided up into three stages:

- ***Stage 1:*** *Validation of the assumption "Aptitude to be classified"*

The objective is to check that the appraiser (or appraisers) is able to classify the products with repeatability and reproducibility. We will determine a numeric variable characterizing the quality of the product.

- ***Stage 2:*** *Realization of trials and classification of results*

We build an experiment matrix according to the selected strategy and we carry out the experiments. The results of each test are then classified compared to a reference sample.

- ***Stage 3:*** *Analysis of the design of experiments*

From the experimental results and the classification carried out, we build a numeric variable allowing the exploitation of the experimental design in a traditional way.

2 Aptitude to qualify the defect

2.1 Principle

The objective of this stage is to check that criterion we seek to evaluate can be qualified or compared in a precise way. The principle of defect evaluation is based on the visual comparison of samples. This comparison is carried out by ranking products from the slightest defect to the most marked. The test consists in validating the reproducibility of the evaluation (or repeatability if carried out by the same appraiser) by presenting a sample of five products to two appraisers (Table 1).

Classification 1	Prod. No.2	Prod. No.5	Prod. No.3	Prod. No.4	Prod. No.1
Classification 2	Prod. No.5	Prod. No.2	Prod. No.3	Prod. No.4	Prod. No.1

Table 1 - Classification of the same sample by two appraisers (from the slightest defected product to the most marked one)

We validate the qualification of the defect by asking one or more people to compare a batch of products several times and classify them according to their perception of quality. If the criterion is perfectly appraisable, the several evaluations would have to give the same result.
To characterize the equivalence of classifications, we calculate a numeric variable representing the quality of this equivalence. We chose Kendall's tau statistic [6].

2.2 Rank correlation Test [7][8]

We are interested in non-parametric correlation tests as we are working with ranks. The aim is to measure the association between two variables (in our case two classifications) and to test their dependence. The theory of ranks performs two well-known correlation tests when two variables are tested: Spearman's Rho (ρ) [9] and Kendall's Tau (τ) [6] [10]. Rho and Tau are defined in [*-1;1*]: they are equal to 1 when the two rankings are identical, they are equal to -1 when the two rankings are completely inverted. The rank correlation coefficient is null when the two variables are independent.
Spearman's Rho is calculated using the rank difference. Kendall's Tau attempts to look at how many times the second ranking is in the same order as the first ranking. Among studies comparing τ and ρ, no obvious differences were found but one can say that "τ approaches normality more rapidly than ρ does" [11] [12] and τ is more interpretable [13]. Moreover τ provides a simple interpretation of the strength of the relationship between two variables. Kendall's Tau also allows to calculate partial correlation based on ranks.
For these reasons we chose Kendall's tau to evaluate the repeatability and reproducibility of several classifications (by the same controller or by two different controllers).
We now briefly detail how to compute τ and test it. We disposed of two observations $(x_1,x_2,\ldots,x_n)$ and $(y_1,y_2,\ldots y_n)$ corresponding to the rankings of objects by two appraisers *X* and *Y* as we wish to qualify the dependence between *X* and *Y*.
After ordering the first ranking, Kendall's Tau attempts to look at how many times the second ranking is in the same order, i.e. in direct order. Each pair of objects of the second ranking is studied. For each pair of compared objects, a score of +1 is given if they are in the direct order and a score of -1 is given if they are in the inverse order. The total of positive scores (also called concordant pairs) and negative scores (also called discordant pairs) is S and τ is calculated with the formula [6]:

$$\tau = \frac{2S}{n(n-1)} = \frac{4P}{n(n-1)} - 1 \qquad \text{[Eq.1]}$$

where *n* is the total number of observed objects and *P* is the number of concordances that is the number of pairs (x_i,x_j) and (y_i,y_j) satisfying either $x_i < x_j$ and $y_j < y_j$, or $x_i > x_j$ and $y_j > y_j$. *P* could be computed by:

$$P = \sum_{i<j} 1_{\{(x_i - x_j)(y_i - y_j) > 0\}} \qquad \text{[Eq. 2]}$$

Hence Kendall's Tau measures the dependence in a very intuitive way as the sign of the product $(x_i-x_j)(y_i-y_j)$ is a characteristic of the correlation between both variables.
In example shown in Table 1, $S = 8$, $P = 9$ and $\tau = 0.8$.
The null hypothesis of independence H_0: $\tau = 0$ could then be tested with the appropriate *p*-value:

- When $n \leq 10$ the exact *p*-value is given by Kendall's tables referencing *p*-values according to the values of *n* and S.
- When n < 10 the standard normal test statistic is:

$$z = \frac{\tau}{\sqrt{\frac{2(2n+5)}{9n(n-1)}}} \qquad \text{[Eq. 3]}$$

Note that when $n < 30$, the *z* approximation is not very accurate and is generally not recommended [10]. Kendall's tables also give significance points of *p*-value for $10 < n < 30$.

In the case of *m* rankings several papers among which [14] propose to compute the average of τ values. Kendall also proposes a coefficient of concordance *W* when more than two rankings are observed [10]. In the case of *m* rankings, Kendall's coefficient of concordance *W* is defined by:

$$W = \frac{12S_d}{m^2(n^3 - n)} \qquad \text{[Eq. 4]}$$

Where S_d is the sum of the squares of the deviations:

$S_d = \sum_{i=1}^{n} R_i^2 - \frac{nm^2(n+1)^2}{4}$ and R_i is the sum of the ranks of the i^{th} ranking.

The coefficient of concordance *W* is equal to 1 when all the classifications are concordant.

To test the significance of an observed value of *W*, Kendall gives tables for $n = 3$ and $m = 2$ to 10, $n = 4$ and $m = 2$ to 6, $n = 5$ and $m = 3$. For other values, an approximation based on Fisher's *z*-distribution [15] could be used and for $n > 7$ one can utilize an approximation based on χ^2 (cf. [6]).

3 Use of the Mann-Whitney variable to obtain a numerical response in the case of an experimental design

3.1 Construction of a numeric variable starting from a classification

In the case of experimental design with a qualitative response, we carried out *N* repetitions for each run in order to constitute an ordered sample according to quality criteria. We suppose a reference sample made up of *m* products. We will discuss the various strategies to constitute this reference sample in paragraph 4. We thus have two samples that it is possible to order according to quality.

We propose to calculate the Mann-Whitney variable traditionally used to test for independence between two sets of variables from these two samples (or also the Wilcoxon variable [6]). It is a non-parametric rank test.

This variable is built in the following way. Considering $(x_1,x_2,\ldots,x_n)$ and $(y_1,y_2,\ldots y_m)$ two ranking samples supposed to come from the same population, with $n \leq m$; the independence is measured by counting the number of couples (x_i, y_j) such that $x_i > y_j$. The total is noted *U*. The *U* variable can vary between 0 to $n \cdot m$. Then:

- $U = 0$ corresponds to the case $y_1 < y_2 < \ldots < y_m < x_1 < x_2 < \ldots < x_n$
- $U = n \cdot m$ corresponds to the case $x_1 < x_2 < \ldots < x_n < y_1 < y_2 < \ldots < y_m$.

If both samples result from the same population, then the variable U has the following characteristics:

$$E(U) = \frac{n \cdot m}{2} \qquad \text{[Eq. 5]}$$

$$Var(U) = \frac{n \cdot m(n + m + 1)}{12} \qquad \text{[Eq. 6]}$$

The variable distribution *U* can be approximated by the normal distribution when *n* and *m* are higher than *8* with relatively good accuracy.

The construction of the *U* variable supposes that all the products are different. In the event of equality (or indecision by the experts)**,** we choose the order of the products randomly. This action is taken into account in the random part of the *U* variable distribution.

3.2 Example

We consider the classification by an appraiser of 5 defects products (tested products) by comparison to 5 products of reference (Figure 1).

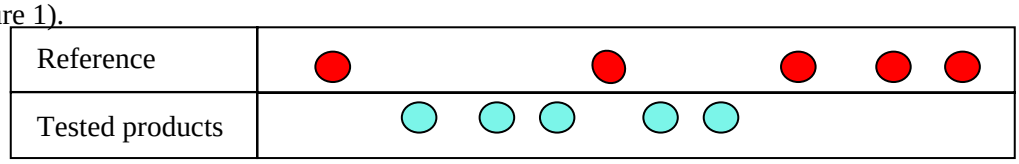

Figure 1 - Classification of the various runs and reference sample

For this classification we calculate the *U* variable (Table 2).

		Rows of the reference				
		1	5	8	9	10
Rows of the test	2	1	0	0	0	0
	3	1	0	0	0	0
	4	1	0	0	0	0
	6	1	1	0	0	0
	7	1	1	0	0	0

U = 7

Table 2 - Calculation of the Mann-Whitney variable

The *U* variable, being calculated for each experiment, allows us to transform a classification into a numeric variable distributed according a roughly normal law when factors studied in the design experiments do not have significant effects. This variable can be seen as a measure of the relative quality of the results of each test compared to the reference sample. We can thus calculate their effects and interactions to determine the contributions of each one of the factors.

3.3 Z test starting from the Mann-Whitney variable

Concerning the *U* variable, the convergence towards a normal law is fast [16], and it is easy to calculate a priori the variance of distribution [Eq 6].
This property of the *U* variables is very interesting because it allows carrying out in a relatively rigorous way a Z test to determine the significant factors by comparing the effect of each factor with the null hypothesis.

4 Examples

4.1 Strategy 1 – Creation of a reference sample

This application is about the thermoforming of an automobile dashboard element. The main critical criteria are aspect criteria. According to the adjustments process, we note the appearance of recurring defects for which descriptors are: burnt folds, cracks. We describe the case of the burnt folds. The same procedure can be use for the second descriptor. An evaluation procedure was defined and validated by a Kendall test on 10 products which gave a p-*value = 0.002213*.
The objective of the company is to find the correct process adjustments which minimize the defects. The process analysis leads to retain three factors:

- Preheating Time [TP]
- Heating Temperature [T]
- Heating Time [TC]

The matrix selected is a complete design 2^3 in order to estimate some interactions. The design of experiments is given in Table 3 and the illustration of the products ranking in Figure 2. For each run, we process 5 products. The appraisers (enterprise experts of control) have to rank all the 40 products coming from the design of experiments and the 5 products of the reference sample.

N°	*TP*	*T*	*TC*	*Test rows*	*Reference rows*	*U*
1	1	1	1	2,3,4,6,7	1,5,8,9,10	7
2	1	1	2	2,4,5,6,8	1,3,7,9,10	10
3	1	2	1	1,2,3,4,6	5,7,8,9,10	1
4	1	2	2	1,2,3,5,6	4,7,8,9,10	2
5	2	1	1	3,5,8,9,10	1,2,4,6,7	20
6	2	1	2	4,7,8,9,10	1,2,3,5,6	23
7	2	2	1	4,5,6,8,10	1,2,3,7,9	18
8	2	2	2	3,5,6,8,9	1,2,4,7,10	16

Table 3 – Design of experiments

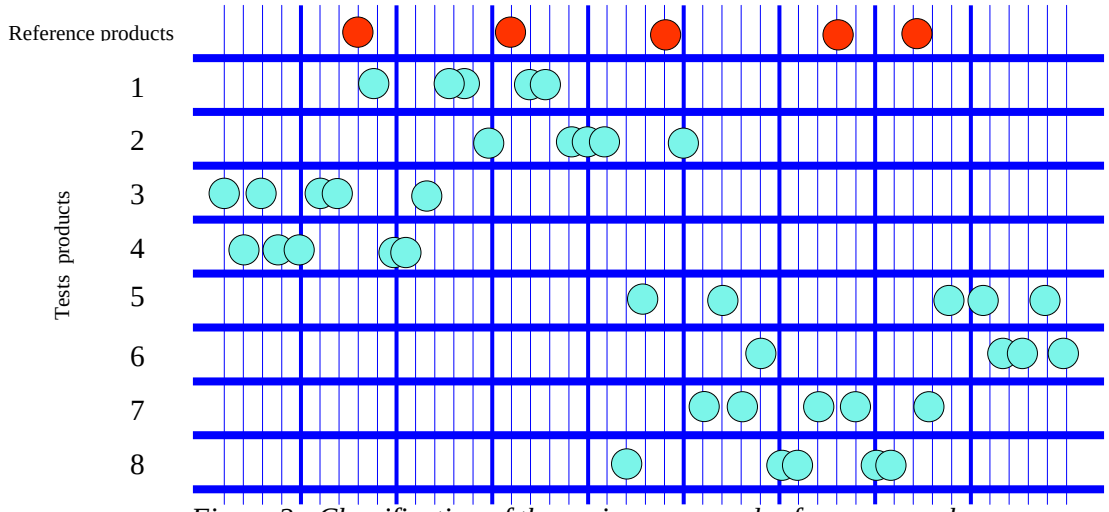

Figure 2 - Classification of the various runs and reference sample

Reference sample choice

The best results will be obtained with a reference sample covering a large scale of the visual evaluation. Indeed, if we choose a sample reference better than all of the design of experiments results, these results will have a high *U* variable. Thus, it will be difficult to determine the effects of each factor. The reference sample can be chosen after the realization of the design of experiments.

Results analysis

The analysis of the experimental design is immediate from the Mann-Whitney variable. Figure 3 shows the effects plot for the factors and interactions.

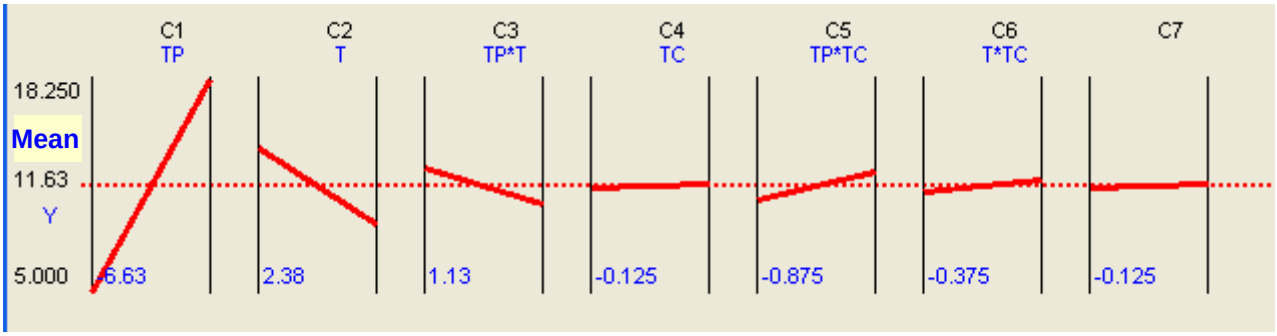

Figure 3 – Effects plot

Z Test on the coefficients

The calculation of the theoretical residual variance is (*n=m=5*):

$$Var(U) = \frac{n \cdot m(n+m+1)}{12} = 22.91 \quad \text{[Eq. 7]}$$

Starting with this variance; we can realize the z test (Table 4).

Term	effect	Coeft	SE coeft	Z	P
TP	13.25	6.625	3.39	3.91	4.5E-05
T	4.75	-2.375	3.39	1.40	0.080
TC	0.25	0.125	3.39	0.07	0.471
TP*T	2.25	1.125	3.39	0.66	0.253
TP*TC	1.75	-0.875	3.39	0.52	0.303
T*TC	0.75	-0.375	3.39	0.22	0.412
T*TP*TC	0.25	0.125	3.39	0.07	0.471

Table 4 – Z Test

Only the first term is really significant, the second term is more critical. The model found with the main effect is:

$$Y = 11.63 + 6.63 * TP - 2.38 * T \quad \text{[Eq. 8]}$$

The effects of the other factors or interaction are not significant.

4.2 Strategy 2 - fictitious reference sample

In strategy 1, the *U* variable is calculated using a comparative evaluation between two samples. In the case of an experimental design with *n* results, it would have led to carrying out C_n^2 comparisons (to make 2 by 2 comparisons). However, with a reference sample, it becomes possible to place the products on a relative scale which is fixed for all the comparisons. If all the results of the experimental design can be evaluated in only one sequence, the problem of the fixed relative scale disappears. We just need to calculate the *U* variable by supposing a random classification. Instead of seeking a uniform reference sample on the range of product variation, we impose a fictitious reference sample which makes it possible "to graduate" the relative position of the products. In the absence of significant effects of the variables in the experimental design, the *U* variables will be randomly distributed, and no factors will be significant.

This strategy was applied to a situation concerning an undesirable noise in a micro engine. At first sight, this noise can easily be measured; however no correlation could be made between various noise measurements and the unpleasant feeling for the ear. The expert has not the capability to evaluate a product on a continuous scale, but is able to rank different products. The factors shows in Table 5 are studied to improve the product quality.

Factors	Name	Level 1	Level 2
grease	Gra	little	With batch
setting	RS	No	Yes
shock on screw	CV	No	Yes
lapping	Gal	Min	Max
ring stuck	Bag	No	Yes
cleanliness	Pro	No	Yes
shock on wheel	CR	No	Yes

Table 5 - Studied Factors

A resolution IV design of experiments was carried out with 16 runs (Table 6) with a response evaluated by an expert. Two micro engines were realized for each configuration of the experimental design.

	GRA	RS	CV	Gal	Bag	Pro	Cr	P1	P2	U Variable
1	Little	Not	Not	Min	Live	Not	Not	3	4	0
2	Little	Not	Not	Max	Boit	Yes	Yes	14	8	4
3	Little	Yes	Yes	Min	Live	Yes	Yes	23	30	12
4	Little	Yes	Yes	Max	Boit	Not	Not	17	9	6
5	Much	Not	Yes	Min	Boit	Not	Yes	26	31	13
6	Much	Not	Yes	Max	Live	Yes	Not	19	18	8
7	Much	Yes	Not	Min	Boit	Yes	Not	13	15	6
8	Much	Yes	Not	Max	Live	Not	Yes	10	16	5
9	Much	Yes	Yes	Max	Boit	Yes	Yes	24	27	11
10	Much	Yes	Yes	Min	Live	Not	Not	11	21	7
11	Much	Not	Not	Max	Boit	Not	Not	7	22	6
12	Much	Not	Not	Min	Live	Yes	Yes	20	25	10
13	Little	Yes	Not	Max	Live	Yes	Not	1	5	1
14	Little	Yes	Not	Min	Boit	Not	Yes	2	6	1
15	Little	Not	Yes	Max	Live	Not	Yes	12	28	8
16	Little	Not	Yes	Min	Boit	Yes	Not	29	32	14

Table 6 - Stamp test with 2 repetitions

The strategy here was not to use a reference sample as in the preceding example, but to classify the 32 micro-engines according to the noise. Starting from this classification, we created a fictitious reference sample uniformly distributed on the whole field covered by the tests. The ranks of the 32 micro-engines are shown in Table 7. In this example, the reference sample contains 7 fictitious engines.

The *U* variable is then calculated by comparing the new ranks of engines of each test to the rank of the fictitious sample reference. In this case, the residual variance is (*n=2;m=7*):

$$Var(U) = \frac{n \cdot m(n+m+1)}{12} = 11.67 \qquad \text{[Eq. 9]}$$

This variance makes it possible to test the level of significance of each factor with a Z test (Table 8).

Row	Rank	Row	Rank	Row	Rank	Row	Rank
1	13-1	11	4-2	21	4-1	31	12-2
2	14-1	12	8-1	22	6-2	32	5-1
3	1-1	13	10-1	23	6-1	33	9-2
4	1-2	14	15-1	24	12-1	34	15-2
5	Ref 1	15	Ref. 3	25	Ref. 5	35	Ref. 7
6	13-2	16	7-1	26	10-2	36	16-1
7	14-2	17	2-1	27	11-2	37	3-2
8	11-1	18	7-2	28	3-1	38	5-2
9	2-2	19	8-2	29	9-1	39	16-2
10	Ref. 2	20	Ref. 4	30	Ref. 6		

Table 7 - Insertion of one fictitious reference

Notation in Table 7: Row 1 – Rank 13-1 means that the best result (1) is found in the trial 13, product of the P1 column.

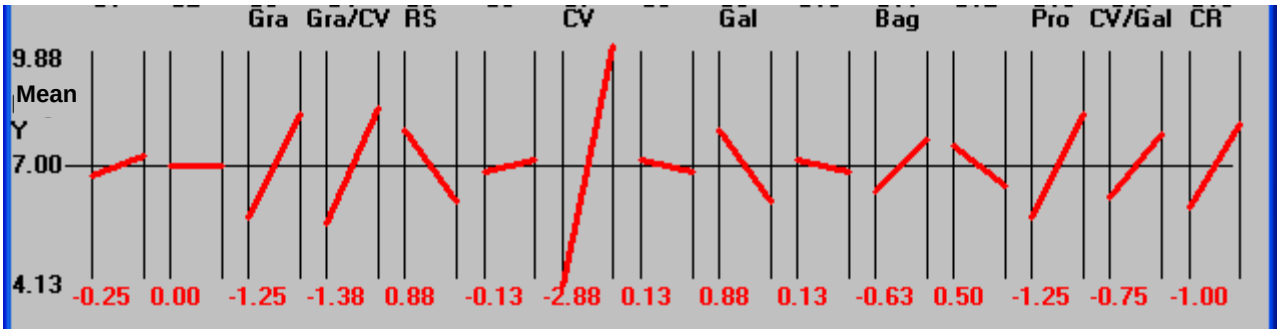

Figure 4 - Result of the plan

Term	effect	Coeft	SE coeft	Z	p
1	0.5	0.25	1.71	0.29	0.385
2	0	0	1.71	0.00	0.500
Gra	2.5	1.25	1.71	1.46	0.072
Gra*CV	2.76	1.38	1.71	1.62	0.053
RS	1.76	0.88	1.71	1.03	0.151
6	0.26	-0.130	1.71	0.15	0.439
CV	5.76	2.880	1.71	3.37	0.000
8	0.26	0.130	1.71	0.15	0.439
Gal	1.76	0.880	1.71	1.03	0.151
10	0.26	0.130	1.71	0.15	0.439
Bag	1.26	-0.630	1.71	0.74	0.230
12	1	0.500	1.71	0.59	0.279
Pro	2.5	-1.250	1.71	1.46	0.072
CV*Gal	1.5	-0.750	1.71	0.88	0.190
CR	2	1.000	1.71	1.17	0.121

Table 8 - Test Z on coefficients

The Z test reveals a very significant factor (CV) and two factors whose level of significance is lower (Gra and Pro). One interaction slightly appears significant. It is easy to identify the active interaction by applying the heredity principle starting from the active factors.

5 Discussion of the method

The aim of this section is to discuss the sampling technique and its effect on the Mann-Whitney variable and on the accuracy of the results of the design of experiments. The Mann-Whitney variable is influenced by:

- Sample size.
- Position of reference sample compared to the position of the experimental results.
- Range of the reference sample compared to the range of the experimental results.

5.1 Sample Size

In the calculation of the Mann-Whitney variable, the reference sample transforms a relative position of an individual into a graduation. Ideally, individuals constituting the reference sample must cover all the range of experimental results uniformly.

The larger the reference sample size is, the more the graduation will be accurate due to the largest range of *U*. However, it is known that *U* variable is biased when sample sizes are unequal [17] [18]. It is thus recommended to have samples of equivalent size.

5.2 Position of the reference sample

The position of the reference sample compared to the range of experimental results can skew the values of *U*. Indeed, if the reference sample is positioned in the extremum ranks, then *U* will be close to *0* or close to *m.n*.

The bias introduced by the reference sample position is partly removed in the effects calculation, since the mean of experimental results (U) is subtracted. Thus, the calculations are relatively robust in spite of a bad centering.

However, the best practice is to choose reference sample centered on the range of the experimental results to limit bias and increase the resolution of the results.

5.3 Range of the reference sample

It is often assumed that heterogeneous variances do not affect nonparametric tests, like the Mann-Whitney test, when sample sizes are equal. In fact, non parametric variables are sensitive to any difference in the shape of distributions, not just in differences of location [8], [19].

We present two extreme situations where U is affected by the homogeneity of the reference sample:

- If the elements of the reference sample are similar (small variance for continued variables), the results of the U variable tend towards 0 or $m.n$ because experimental results are either higher ranked or lower ranked (there is no covering between the individual samples), as the effects of the result are very discretized (few possible values).
- Conversely if the reference sample covers a large range of experimental results (large variance for continued variables), the U values tend towards the average $U = m.m/2$ and then the effects tend towards 0.

These situations are voluntarily extreme in order to show the importance of choosing a reference sample covering the range of experimental results while avoiding retaining the extreme elements.

6 Conclusion

In this article, we showed that it is possible to use an experimental design on non measurable responses by using a rank classification.

Among the interests of the method, we can note the following arguments:

- Possibility of using experimental designs with an organoleptic response.
- Simple method exploiting the standard statistical tools, which allows an implementation without using specific software and simple interpretation of the results.
- Possibility of using the Z test to determine the significant factors.

However, to use this technique, it is necessary to meet the following three conditions:

- The criteria can be classified in order of quality or of criticality. It is recommended to clearly define the classification criteria: defect intensity, numbers, localization or semantic scale of classification.
- The choice of the factor levels must generate variations interpretable or discernible by appraisers. If not, U values will be close to $U = \frac{m \cdot n}{2}$ and all effects will be null or negligible.
- The reference sample must be chosen with a uniform distribution on the whole field covered by the tests. The size of the reference sample should be close to the size of the experimental sample. It is possibly to choice a fictitious reference sample.

References

[1] ISO 13299:2003 Sensory analysis -- Methodology -- General guidance for establishing a sensory profile, 2003.

[2] F. Depledt, Evaluation sensorielle : Manuel méthodologique, Tech. & Doc., Lavoisier, 1998.

[3] M. C. Gacula, Descriptive sensory analysis in practice, Wiley-Blackwell, 1997.

[4] M.C. Meilgaard, G.V. Civille and B.T. Carr, Sensory evaluation techniques, 3rd ed. CRC. CRC Press, London, 1999.

[5] M. O'Mahony, Sensory evaluation of food, Statistical methods and procedures – Marcel Dekker – 1986.

[6] M.G. Kendall, A New Measure of Rank Correlation, Biometrika, 30, pp. 81-93, 1938.

[7] H. Jaroslav, S. Zbynek and K. Pranab, Theory of Rank Tests, 2nd ed, Academic Press, 1999.

[8] E.L. Lehmann, Nonparametrics: Statistical methods based on ranks, San Francisco: Holden-Day, 1975.

[9] C. Spearman, The Proof and Measurement of Association Between Two Things, American Journal of Psychology, 15, pp. 72-101, 1904.

[10] M. Kendall and J.D. Gibbons, Rank correlation methods, 5th ed, Ed. Edward Arnold, 1990.

[11] S. Yue, P. Pilon and G. Cavadias, Power of the Mann-Kendall and Spearman's rho tests for detecting monotonic trends in hydrological series, Journal of Hydrology, 259, pp. 254-271, 2002.

[12] W.W. Daniel, Applied Nonparametric Statistics, Boston: PWS-Kent, 2nd edition, 1990.

[13] G.E. Noether, Why Kendall Tau?, The best of teaching statistics, pp. 41-43, 1986.

[14] M. Alvo and P. Cabilio, Average rank correlation statistics in the presence of ties, Communications in Statistics – Theory and Methods, 14, 1985.

[15] R.A. Fisher, Statistical methods for research workers, Oliver & Boyd, Edinburgh, 1934.

[16] Ph. Caperaa and B. Van Cutsem, Méthodes et modèles en statistiques non paramétriques Exposé fondamental, Presse de l'université Laval, DUNOD, 1988.

[17] D.W. Zimmerman, A note on homogeneity of variance of scores and ranks, Journal of Experimental Education, vol. 64, pp. 351-362, 1996.

[18] D.W. Zimmerman and B.D. Zumbo, Rank transformations and the power of the Student t test and Welch t' test for non-normal populations with unequal variances, Canadian Journal of Experimental Psychology, vol. 47, pp. 523-539, 1993.

[19] R.H. Randles and D.A. Wolfe, Introduction to the theory of nonparametric statistics. New York: Wiley, 1979.